Problem Statement

The dataset is used to train the machine learning model so that it allows us to predict whether a client is able to repay their loan based on the client's background and credit history.

Findings from data exploration

1. Gender Diagram: There seems to be much more female than male in this dataset
2. Owns a car diagram: There are more people not owning a car than people owning a car
3. Owns a flat diagram: Majority of the people in the dataset do own a flat
4. Contract type diagram: Cash loans seems to be the more popular choice as compared to revolving loans
5. Number of children diagram: The number of people not having any children is way higher than those who owns at least 1 children
6. Education Type: majority of the people here completed at least secondary school
7. Majority of the people in the dataset are married couples.
8. Majority of them stays in houses/apartment
9. Most of them are working, none of them are unemployed or studying
10. The married couples are all staying by themselves
11. Majority of them are laborers with sales staff being the second most popular job
12. The columns AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE all have high correlation.
13. There are way too many TARGET = 0 as compared to TARGET = 1 which might skew the model into having a very high probability of predicting 0 as compared to 1

Machine learning model

The logistic regression model is chosen as the model to classify the targets into two classes, 0 and 1. Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. Hence for this case, the likelihood of a client being able to repay its loan on time will be determined by TARGET = 1 class and the unlikelihood of a client being able to repay its loan on time will be determined by TARGET = 0 class. 0 and 1 being the target of this logistic regression model are known as binary dependent variables.

Model accuracy

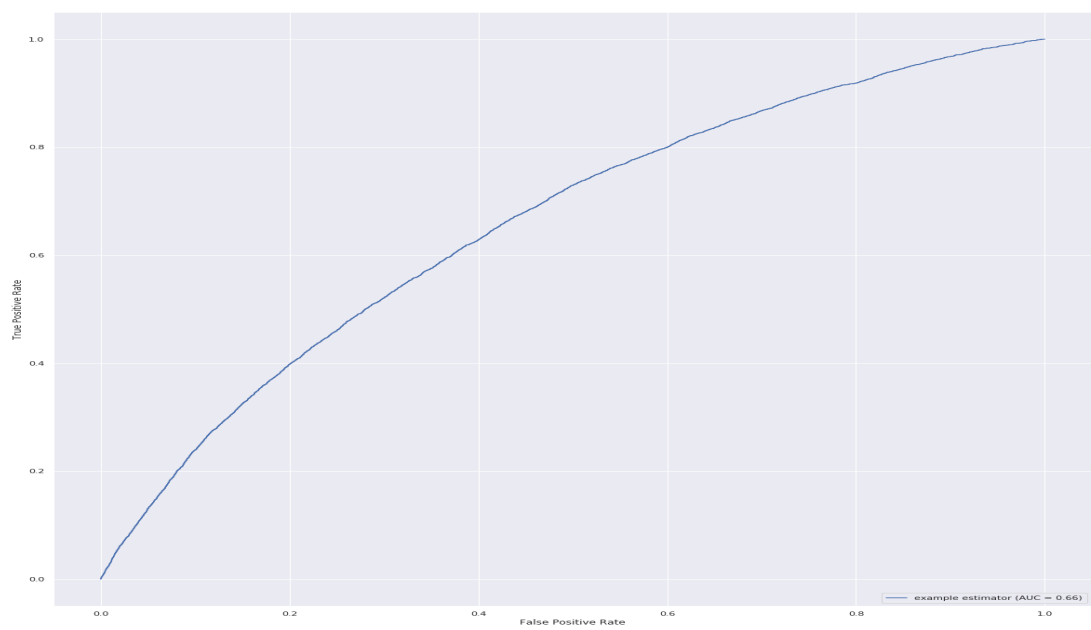
Before hyperparameter tuning:

In the initial setup of the model, the non-object datatype is scaled using min-max scaler and the object datatype such as gender are being one-hot encoded as seen in a screenshotted portion of the table below:

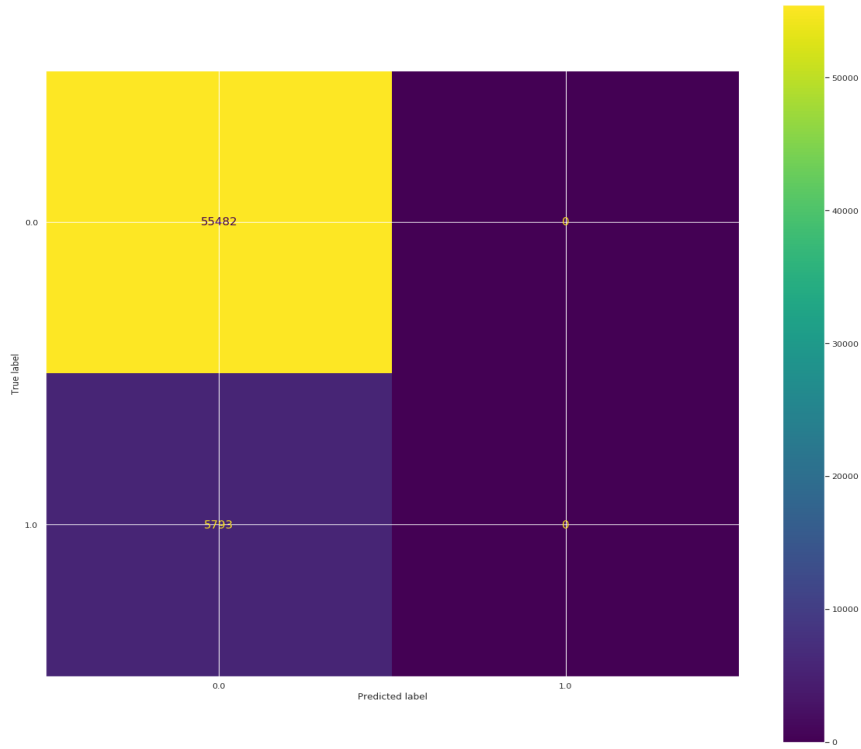
CODE_GENDER_F	CODE_GENDER_M	C
0	1	
1	0	
0	1	
1	0	
0	1	

The model is then split into 70% train dataset and 30% test set and trained onto the logistic regression model with no parameters set, using the default parameters.

The model then generated an accuracy of 90.55% and an AUC score of 0.66 as seen in the attached diagram below:



Using the confusion matrix to identify the predicted values against the true values, I realize that the model predicted all the dataset in the test dataset to the 0 class and none were the 1 class. Hence, I decided to count the number of class 0 data and class 1 data in the TARGET column, realizing that there were almost x10 class 0 data as compared to class 1 data.



After hyperparameter tuning:
After Grid searching with the parameters as shown:

```
search_parameters = {
    'penalty' : ['l1','l2'],
    'class_weight':      ("balanced", None),
    'solver' : ['lbfgs','saga']
}
```

The results of each parameter fitted into the logistic regression model did not produce any results that are much better than the default parameters. The best model selected from this Grid Search, LogisticRegression(penalty='l1', solver='saga'), produced similar results.

Limitations

The data might consist of really old data which causes the dataset to be non-relevant if we want to use it to predict the future. The dataframe consists of too many 0 labels in the TARGET column as compared to 1 labels in the TARGET column making the training very unfair and skewed towards predicting 0. This can be seen from the confusion matrix as to none of the data predicted were label 1. The dataset also consists of many null values which was being filled with mean values of the columns, making the features to this model inaccurate.

Conclusion

As the AUC score is still very low, we can make use of neural network models such as CNN models or even RNN models to make better predictions of this model. The dataset can also be improved by adding the application date of each individual candidate as well as add more class 1 labels into the dataset to achieve a fairness between the 0 and 1 labels of the TARGET column. All in all, the dataset could not achieve a better accuracy score due to the nature of the skewed dataset. Hence, a technique of improving the model could be adding weights into the classes 0 and 1 labels such that the class 0 labels have a smaller weight as compared to the class 1 labels and so after applying neural network training on it, the weights that are multiplied will be relevant with the number of data that appears in the dataset preventing the biasness in the machine learning model.